Depth Uncertainty in Vision Transformers (ViT)

Abhinav Joshi
20211261

Shrey Bhatt
20111060

Akash Halayyanavar
20111006

Saksham Jain
20111053

Abstract

Vision transformers (ViTs) have shown successful applications in computer vision tasks recently. However, a significant drawback of such deep learning models is their inability to provide robust uncertainty estimates. In this work, we explore the idea of depth uncertainty in vision transformers and report our findings in detail. We combine ViT encoder layers with a probabilistic model for estimating uncertainty in depths. We further show an empirical comparison for studying the effects in detail.

1 Introduction

Transformer architectures have gained colossal research interest due to their success in the field of natural language processing (NLP). Some of the famous works include BERT [7] and GPT [4] models. Recently, Transformers have drawn much research attention in various fields of computer vision tasks such as image recognition [8, 27], object detection [5, 33], and image processing [6]. Some of the recent works [15] have shown a successful implementation of transformers to vision tasks [24] with competitive performance compared to Convolution neural networks [17, 26].

Compared to CNNs that extract features by stacking multiple convolution layers using multiple kernels, Vision transformers [8] takes advantage of the self-attention mechanism [29] to capture spatial patterns and non-local dependencies. The ability of ViTs to capture rich global information without the need for layer-wise local feature extractions, as in CNNs, makes them better in terms of performance and computations. However, recent progress of these architectures is largely driven by training models with multiple layers stacked together and using the techniques like residual learning [17] for better performance. This not only limits the use of such networks to a high computation environment, but also makes them harder to train for deep architectures.

Another drawback of such networks is the inability of such networks to provide robust uncertainty estimates. Deep models are often overconfident, even when their predictions are incorrect [23, 1]. Including uncertainty estimation in the model predictions for such networks often lead to an extra burden of computation [9, 18, 13, 21]. For solving the computation issue present in these models, recent methods like DUNs [2] are proposed, which perform probabilistic reasoning over neural networks' depth. Each depth corresponds to a subnetwork that shares weights, and further, the extracted features from these layers are combined via marginalization, yielding the model uncertainty.

Taking inspiration from DUNs, in this work, we explore the use of Vision transformers in an uncertain depth environment. We propose an architecture that combines the Vision transformer encoder and a probabilistic model for depth uncertainty. We study the effect of using depth as uncertainty in Vision transformer and experiment on multiple datasets for comparing ViT with our architecture. The probabilistic model exploits the overparameterization in Vision transformers by marginalizing the features over depth. The use of uncertainty over depth not only helps in regularizing the model but also helps in learning rich representations of the data.

2 Prior Works

Neural networks have been widely studied for uncertainty estimation as Bayesian Neural Networks by placing distributions over weights and translating the weight uncertainty to prediction uncertainty. For faster uncertainty estimation, dropout techniques like [11, 22] have also been employed. Non-Bayesian approaches with high computation cost like ensembles [20] use predictions from multiple models and try to estimate the uncertainty over the predictions. Deep ensembles can also be interpreted as approximate BMA [30]. There are methods which measures the Uncertainty of Predictions in Deep Neural Networks with Variational Inference by representing the a posteriori uncertainty of the network parameters per network layer results in decreasing computational cost compared to a non-Bayesian network [25]. DUNs use a categorical distribution to model depth, do not require sampling for evaluation of the training objective or making predictions, and can be applied to a wider range of Neural Network architectures.
Transformer is used primarily in the field of natural language processing (NLP). It also has applications in tasks such as video understanding. Like recurrent neural networks (RNNs), transformers are designed to handle sequential input data, such as natural language, for tasks such as translation and text summarization. Transformers can also be used for image classification [3]. Visual transformers usually view an image as a sequence of patches while they ignore the intrinsic structure information inside each patch by Transformer-iN-Transformer (TNT) model [16]. Deeper transformer networks for image classification are also available that optimizes the performance of Image transformers [28]. Transformers require minimal inductive biases for their design and are naturally suited as set-functions and demonstrates excellent scalability to very large capacity networks and huge datasets [19].

3 Proposed Architecture

Vision transformers have gained much attention in recent years. The primary idea in vision transformers is to use patches of images as a sequence, and further use them as an input to the transformer architecture employing attention mechanism. They claim to provide better long-term feature retention than other sequential models like LSTM, RNN. Though being powerful architectures, which have shown huge success on various datasets, the ability to train and use them remains limited to a high computation environment. We explore the possibility of using them in low computation environment by combining them with the Depth Uncertainty Networks. The Figure 1 shows a detailed diagram of our architecture.

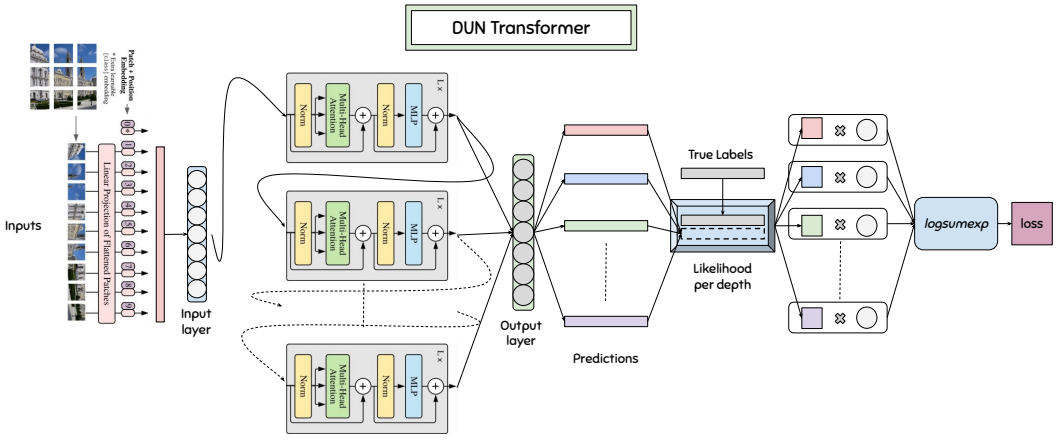

Figure 1: Depth Uncertainty in Vision Transformers

Our architecture is composed of two models, Vision transformer encoders for feature extraction and probabilistic model for depth uncertainty estimation. We take an inspiration from DUNs and

combine a similar architecture to the Vision transformer encoders. We further study the effect of combining the DUNs to ViT with the help of empirical comparison over the Vision transformers.

Our probabilistic model predicts the uncertainty over depths by treating depth as the random variable and network weights as the learnable parameter. A categorical prior is placed over the depth. Using this approach, we were able to compute the posterior predictive mean and variance using the exact posterior in a single forward pass. Both the network weights and the posterior over depth can be obtained by optimising the Marginal Log-Likelihood using the Stochastic gradient variational inference algorithm. For this purpose, a surrogate categorical distribution over depth is used. We use the strategy of variational inference to train the entire model, where an ELBO objective is used to optimize the network parameters and is given as,

$$\log p(\mathcal{D};\theta) \geq \mathcal{L}(\alpha,\theta) = \sum_{n=1}^{N} \mathbb{E}_{q_\alpha(d)} \left[\log p\left(y^{(n)} \mid x^{(n)}, d; \theta\right)\right] - \mathrm{KL}\left(q_\alpha(d) \| p_\beta(d)\right)$$

where,
d : Depth; θ : Network weights; α : Variational parameters; $\mathcal{D} = \{x^{(n)}, y^{(n)}\}_{n=1}^{N}$ is the dataset

Using the gradients of the ELBO, the variational parameters and the network weights can be optimized simultaneously. Since both the variational and true posteriors are categorical, exact inference can be made. The posterior predictive distribution is obtained by marginalising the depth with the variational posterior,

$$p\left(y_* \mid x_*, \mathcal{D}; \theta\right) = \sum_{i=0}^{D} p\left(y_* \mid x_*, d=i; \theta\right)\ q_\alpha(d=i)$$

4 Methodology

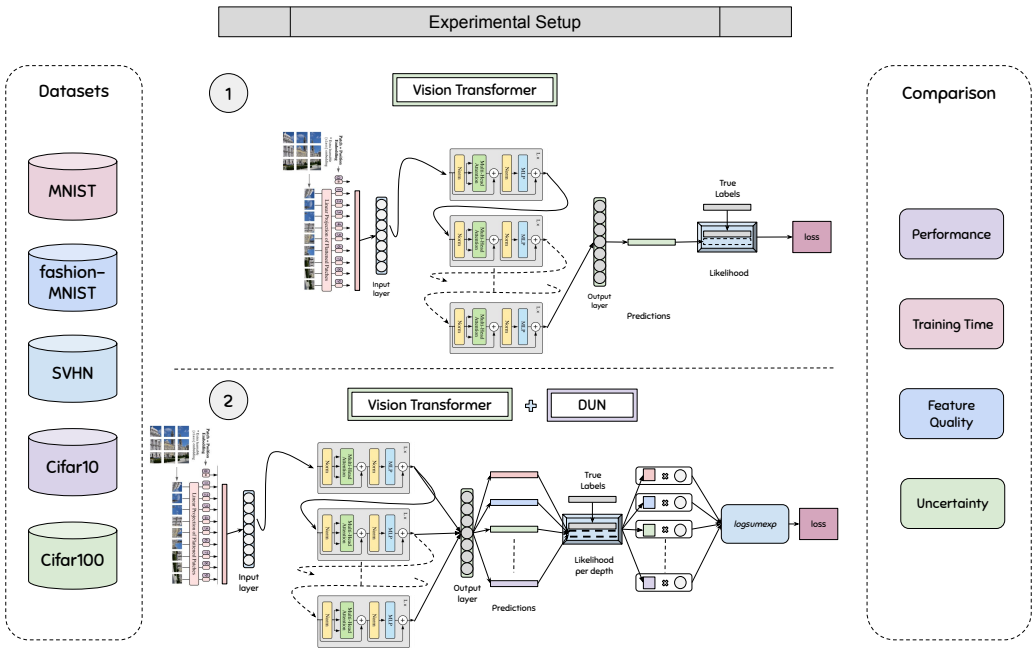

Figure 2: Experimental setup for comparing ViT with ViT + DUN.

For a detailed study of effect of DUNs over ViT, we design our experiments for comparing ViT with ViT + DUNs. We explore this effect with the help of following different measures: Performance,

Effective Feature Extraction, Training/Convergence time, Uncertainty estimation and Regularization effect. For measuring these qunatities, we performed experiments over few datasets like MNIST, CIFAR-10, Fashion MNIST, SVHN. Figure 2 gives an overview of our experimental setup.[1]

Performance: For measuring performance of the model, we analyzed the accuracy and test loss of model over the epochs. Depicted in the image 3 are the plots of same over 25 epochs for MNIST dataset. As observed in the plots, although the ViT + DUN model (shown in orange color) starts off with a higher loss, it quickly converges over to the similar value of loss and accuracy as that of ViT over the epochs. As observed in the plots of MNIST, which is a dataset with relatively simpler features. As observed, after some epochs the ViT+DUN model performs almost similar to ViT in terms of both test loss and accuracy. Both obtain an approximate accuracy of 97-98% with slights fluctuations at the end. Both the models were trained with final depth=3. The accuracy comparison of both models over different datasets are highlighted in the table 1.

Dataset	ViT - Accuracy	ViT + DUN - Accuracy
MNIST	98.05%	97.81%
CIFAR10	56.46%	55.63%
SVHN	80.19%	79.85%
Fashion MNIST	87.83%	87.75%

Table 1: Accuracy over datasets

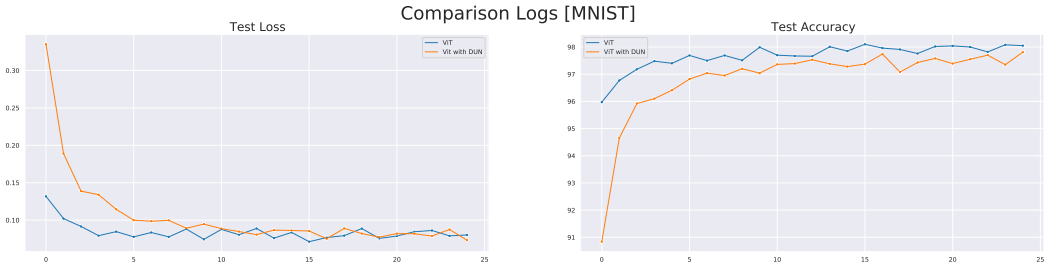

Figure 3: Test Loss and Accuracy over MNIST

Quality of Feature Extraction: For analyzing the quality of features extracted by the different layers of the Vision transformer, we extract activation vectors from each layer of ViT and further plot a t-SNE over them. The t-SNE plot of MNIST dataset over activation vectors from different depths is in Figure 4. The proposed architecture is able to segregate MNIST classes starting from depth one due to the simplicity of the dataset. We further test the feature extraction in the SVHN dataset and report our finding in Figure 5. SVHN being a more complex dataset, shows relatively less segregated clusters for initial depths and starts improving as the number of depths increases.

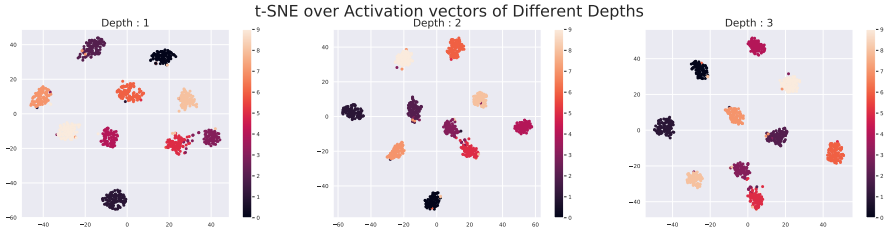

Figure 4: T-SNE of DUN predictions over MNIST

[1]Our experimental setup can be replicated using the python notebook: `https://tinyurl.com/ymdpepru`

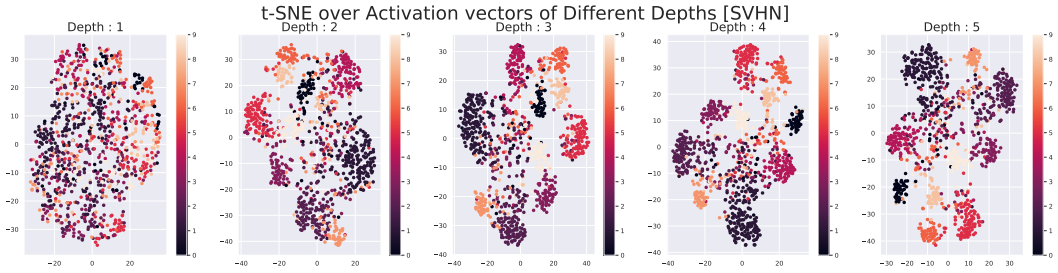

Figure 5: T-SNE of DUN predictions over SVHN

Training/Convergence Time: We further compare both the models using convergence rate and required training time. The Table 2 highlights the time taken by both models over different datasets. As clearly observed, both models have very less significant difference of training time. This is due to the nature of DUNs that only require a single forward pass similar to their ViT counterpart. This makes it less costlier for uncertainty estimation than some methods that require multiple forward pass for parameter estimation.

Dataset	ViT - Training Time	ViT + DUN - Training Time
MNIST	627 s	692 s
CIFAR10	1824.29 s	1838.47 s
SVHN	1083.33 s	1056.10 s
Fashion MNIST	615.16 s	621.39 s

Table 2: Training time over datasets (per epoch)

Uncertainty Estimation: For uncertainty estimation, we perform predictions over rotated images. Since image datasets like SVHN, MNIST and Fashion MNIST do not lose features on slightly lower rotation angles but they might be unrecognizable on higher angles, the evaluation of these models on rotated images could be used to check if they have learnt the shapes of the images in the dataset. The performance of ViT and ViT + DUN over rotated MNIST dataset is as depicted in the figure 6. As shown, the DUN model (in green) performs slightly better than ViT (in black) over angles from 50 to 150 degrees and performs similar for other angles. We also performed the same experiment over SVHN dataset (figure 7) as they also possess housing number based feature representation. The performance of both the models in this case, was quiet competitive. However, in this case, the accuracy of both models start quickly worsening over at early rotation angles as compared to MNIST, possibly because of slightly more complex data and background shapes also being present in the images.

Regularizing Effect: Another important finding that we observed empirically is the regularizing effect of using a probabilistic model. The test loss plot over epochs, clearly depicts the overfitting of ViT without DUNs. A possible reason for this could be the use of marginal prediction of DUNs during training that allowed it to produce relatively stable predictions. To observe this result more carefully, we experimented over a relatively more straightforward Fashion MNIST dataset and let both the models run over it for 40 epochs with depth=10. As expected, due to the large number of parameters used, the test loss of ViT architecture started increasing after 20 epochs. In contrast, the ViT+DUN tends not to overfit even for a high number of parameters. DUNs exhibit a regularizing effect over ViT encoder layers and make it robust towards overparameterization. However, the accuracy of both models remained intact. A possible reason is that although the softmax activation of ViT started producing worse values, the predictions were still correct.

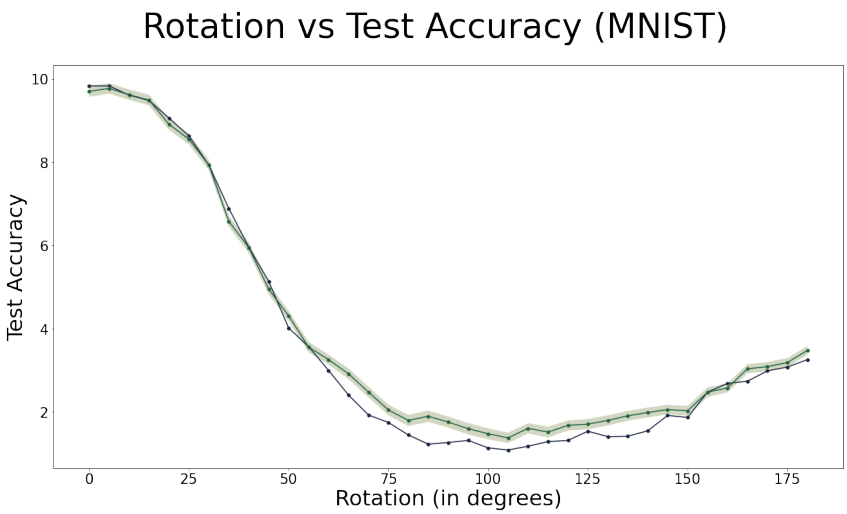

Figure 6: Performance over rotated MNIST

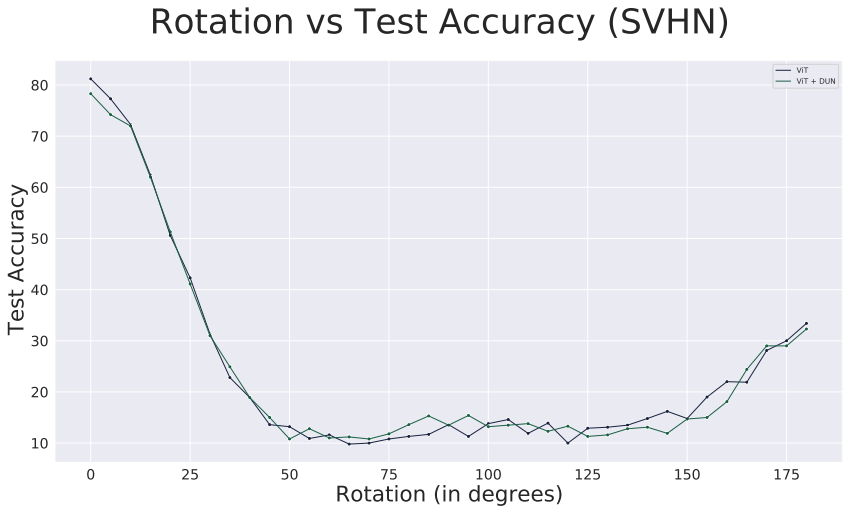

Figure 7: Performance over rotated SVHN

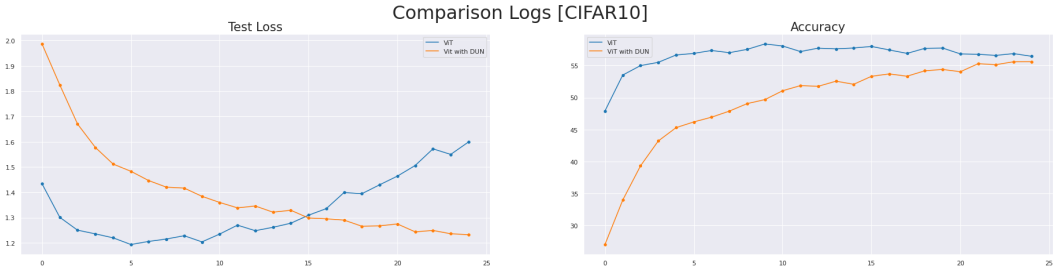

Figure 8: Test Loss and Accuracy over CIFAR-10

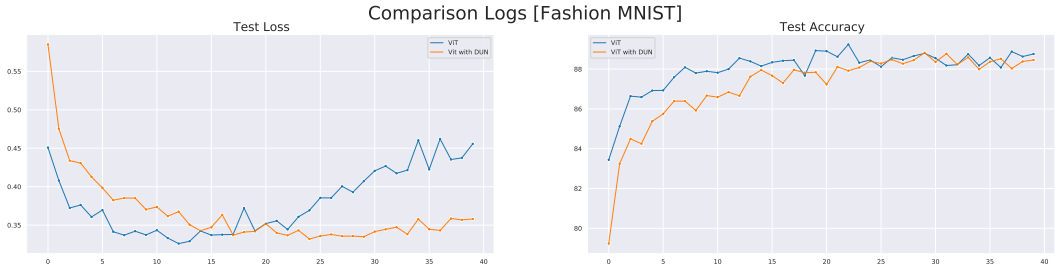

Figure 9: Test Loss and Accuracy over Fashion-MNIST

5 Things we Learnt

In this section, we give a brief overview of methods that we learned during this project work. We divide our learnings into phases; wherein the first phase, we explored various methods of uncertainty estimation in deep learning models. Some of the famous work which we went through include Bayesian Neural Networks [14], Bayesian model averaging [10], Deep Ensembles[9] and MC Dropout [12]. We further studied the Depth uncertainty networks in detail and experimented with their implementation on other regression datasets like house sales price prediction[2].

In the second phase of this work, we tried to expand the use of DUNs to transformer encoder architectures. We explored the Vision Transformer and its related work in the field of computer vision. We also went through few works that cover uncertainty estimation in transformers for natural language processing tasks like [31]. We further combined the idea of depth uncertainty with vision transformers and made an empirical comparison for studying the effects of DUN over vision transformers.

6 Conclusion and Future Directions

In this work, we explored the idea of Depth uncertainty in Vision Transformers. We proposed an architecture that combines Vision transformer encoder layers with a probability model and estimates the depth uncertainty, which further is translated to the prediction uncertainty using marginalization. We studied the effects of depth uncertainty over vision transformers empirically and reported our findings in detail.

Some of the future directions for our work include testing transformer models in a sequential dataset environment. We explored the use of depth uncertainty in vision transformers using some popular image datasets. Transformers, being suitable architectures for sequential datasets, can be explored in a similar setting for estimating depth uncertainty. Moreover, one can always use pre-trained transformers and study the effect of including a probabilistic model for depth estimation. Another possible line of work consists of the study of deep architectures like DeepViT[32] and using depth uncertainty networks for optimum neural architectural search (NAS).

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016.

[2] Javier Antorán, James Urquhart Allingham, and José Miguel Hernández-Lobato. Depth uncertainty in neural networks, 2020.

[3] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. *arXiv preprint arXiv:2103.14586*, 2021.

[2]House Sales Price prediction dataset link: `https://www.kaggle.com/rajeshdgr8/housetrain`

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer, 2020.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[9] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective, 2020.

[10] Tiago M. Fragoso, Wesley Bertoli, and Francisco Louzada. Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, 86(1):1–28, Dec 2017.

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016.

[13] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns, 2018.

[14] Ethan Goan and Clinton Fookes. Bayesian neural networks: An introduction and survey. *Lecture Notes in Mathematics*, page 45–87, 2020.

[15] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A survey on visual transformer, 2021.

[16] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[18] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get m for free, 2017.

[19] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021.

[20] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

[21] Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning, 2019.

[22] Eric Nalisnick, José Miguel Hernández-Lobato, and Padhraic Smyth. Dropout as a structured shrinkage prior. In *International Conference on Machine Learning*, pages 4712–4722. PMLR, 2019.

[23] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, 2015.

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015.

[25] Jan Steinbrener, Konstantin Posch, and Jürgen Pilz. Measuring the uncertainty of predictions in deep neural networks with variational inference. *Sensors*, 20(21):6011, 2020.

[26] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[27] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers distillation through attention, 2021.

[28] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. *arXiv preprint arXiv:2103.17239*, 2021.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[30] Andrew Gordon Wilson. The case for bayesian deep learning. *arXiv preprint arXiv:2001.10995*, 2020.

[31] Boyang Xue, Jianwei Yu, Junhao Xu, Shansong Liu, Shoukang Hu, Zi Ye, Mengzhe Geng, Xunying Liu, and Helen Meng. Bayesian transformer language models for speech recognition, 2021.

[32] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer, 2021.

[33] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection, 2021.